GENERALIZATION ERROR ANALYSIS FOR SPARSE MIXTURE-OF-EXPERTS: A PRELIMINARY STUDY

Jinze Zhao[1], **Peihao Wang**[1], **Zhangyang Wang**[1]
[1]Department of Electrical and Computer Engineering, UT Austin

ABSTRACT

Mixture-of-Experts (MoE) represents an ensemble methodology that amalgamates predictions from several specialized sub-models (referred to as experts). This fusion is accomplished through a router mechanism, dynamically assigning weights to each expert's contribution based on the input data. Conventional MoE mechanisms select all available experts, incurring substantial computational costs. In contrast, Sparse Mixture-of-Experts (Sparse MoE) selectively engages only a limited number, or even just one expert, significantly reducing computation overhead while empirically preserving, and sometimes even enhancing, performance. Despite its wide-ranging applications and these advantageous characteristics, MoE's theoretical underpinnings have remained elusive. In this paper, we embark on an exploration of Sparse MoE's generalization error concerning various critical factors. Specifically, we investigate the impact of the number of data samples, the total number of experts, the sparsity in expert selection, the complexity of the routing mechanism, and the complexity of individual experts. Our analysis sheds light on *how **sparsity** contributes to the MoE's generalization*, offering insights from the perspective of classical learning theory.

1 INTRODUCTION

Sparse Mixture of Experts (SMoE) (Shazeer et al., 2017), has demonstrated remarkable potential in the realm of neural networks, particularly in the context of scaling model size without incurring significant computational overhead. SMoE achieves this by ingeniously partitioning neural networks into smaller, more focused networks known as *expert* networks, which are then selectively and sparsely combined using a data-dependent neural network referred to as the *router*. This approach results in a substantial increase in the total number of parameters while keeping computational costs relatively stable, owing to the fact that only a small subset of experts is chosen for each data point. Furthermore, SMoE is promising to improve model generalization, particularly when dealing with multi-domain data, as it allows a group of experts to learn collaboratively and generalize to unseen domains compositionally. However, despite these compelling advantages, our theoretical understanding of MoE remains elusive. It is often counterintuitive that increasing model size would enhance generalization, as conventional wisdom suggests the opposite effect.

To address this paradox, this work presents the generalization bound for the Sparse MoE model, by drawing inspiration from classical learning theory concepts (Natarajan, 1989; Vapnik, 1995; Shalev-Shwartz & Ben-David, 2022). We assert that our generalization bound is **model-agnostic**, offering **versatility** and **applicability** across various expert/router-based models when the upper bounds for the complexity metrics of these base models are incorporated. Our main results can be informally stated as:

- We delivered a **generalization error bound** for SMoE, which depends on (only) its high-level structure hyperparameters. Our generalization error bound can be applied to generic SMoE model, regardless of the expert or router implementation.
- Our generalization bound is in particular **sparsity-aware**. More specifically, Theorem 1 shows that the generalization error scales with the 'sparsity pattern' by $O\left(\sqrt{k(1+\log(\frac{T}{k}))}\right)$ where T is the number of all experts and k is the number of selected experts, meaning the generalization error bound will grow tighter once we select less

experts. The analysis sheds light on *how **sparsity** contributes to the MoE's generalization even at increasing total model sizes*, compared to conventional MoEs selecting all available experts.

1.1 Related Work

Mixture of Experts: The Mixture of Experts (MoE) model, originally introduced by Jacobs et al. (1991) and later extended by Jordan & Jacobs (1994), is a framework that integrates multiple specialized sub-models through a router network. This enables the adaptive weighting of their contributions based on the input data distribution, thereby enhancing the overall predictive performance. Recent advancements in the field, such as the works by Shazeer et al. (2017) and Eigen et al. (2013), have harnessed this concept within the realm of Deep Learning. These efforts introduced 'routing sparsity', which significantly reduces computational costs and enables the scaling of models to sizes with trillions of parameters. More recently, Fedus et al. (2022) and Jiang et al. (2024) pushed the boundaries of sparsity in MoE models by aggressively routing input data to sparse experts while paradoxically preserving performance and scaling up model size in language modeling.

Theoretical Understanding of MoEs: Recent work (Chen et al., 2022) formally studied how the SMoE layer reduces the training error better than using a single expert, and why such mixture model will not collapse into a single model. Importantly, when training an SMoE layer based on the data generated from the "mixture of class" distribution using gradient descent, the authors proved that each expert of the SMoE model will be specialized to a specific portion of the data (at least one cluster), and meanwhile the router can learn the cluster-center features and route the input data to the right experts. More recently, a series of work (Nguyen et al., 2023d;c;a;b) tried to establish the convergence rates of density estimation and parameter estimation of MoE by defining Voronoi-based losses which describes the interaction between gating function and experts, and explained why Top-1 gating can lead to faster convergence rates of parameter estimation over other gating mechanisms. However, we opt to formally study the *generalization* benefit of Sparsity in MoE modeling, under learning theory perspectives.

2 Preliminaries

Setup: We mainly focus on *binary classification* for its convenience of analysis. Consider a distribution: $\mathcal{D} \sim \mathbb{P}(\boldsymbol{x}, y)$, where $\boldsymbol{x} \in \mathbb{R}^d$ is a the raw feature vector, and $y \in \{+1, -1\}$ is the corresponding binary label. Same with the classical learning settings, we assume our dataset $\mathcal{S}$ consists of m i.i.d. samples, i.e.,

$$\mathcal{S} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_m, y_m)\} \overset{i.i.d.}{\sim} \mathcal{D}^m \tag{1}$$

Given a learner/classifier $f : \mathbb{R}^d \to \{+1, -1\}$, we define the empirical loss over training set to be $\mathcal{S}$: $L_{\mathcal{S}}(f) = \frac{1}{m}\sum_{i=1}^{m} \ell(f(\boldsymbol{x}_i), y_i)$, and the truth error evaluated over the whole data domain: $L_{\mathcal{D}}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\ell(f(\boldsymbol{x}), y)$, where $\ell : \{+1, -1\}^2 \to \mathbb{R}$ is the loss function. We define a class of functions as the hypothesis space $\mathcal{F}$. At the training stage, our learning algorithm searches for a learner $f \in \mathcal{F}$ which minimizes the empirical loss, i.e., running Empirical Risk Minimization (ERM):

$$\text{ERM}(\mathcal{S}) = \arg\min_{f\in\mathcal{F}} L_{\mathcal{S}}(f), \tag{2}$$

At the testing stage, we compute $L_{\mathcal{D}}(\text{ERM}(\mathcal{S}))$ to assess the model performance. The generalization error of hypothesis $f \in \mathcal{F}$ is:

$$\epsilon(f, \mathcal{S}) = |L_{\mathcal{S}}(f) - L_{\mathcal{D}}(f)| \tag{3}$$

2.1 Notations for Sparse Mixture-of-Experts

We define the (simplified) SMoE (Shazeer et al., 2017) model $f : \mathbb{R}^d \to \mathbb{R}$ as below:

$$f(\boldsymbol{x}) = \sum_{j=1}^{T} a(\boldsymbol{x})_j h_j(\boldsymbol{x}) \quad \text{subject to} \sum_{j=1}^{T} \mathbb{1}\{a(\boldsymbol{x})_j \neq 0\} = k \quad \forall \boldsymbol{x} \in \mathbb{R}^d, \tag{4}$$

where $a(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^T$ is the router function satisfying $\|a(\boldsymbol{x})\|_0 = k$, and $h_j(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ is a expert learner for every $j = 1, \cdots, T$. Intuitively, the entire SMoE model contains T expert learners, thus, its capacity is T times larger as one expert learner. However, at each forward pass, only k experts will be activated for each data sample, leading to comparably low inference costs as activating all experts.

In this paper, we consider the router function proposed by Shazeer et al. (2017), which first chooses the k output logits, sets the remainders to zeros, and applies softmax to normalize the chosen entries:

$$a(\boldsymbol{x})_j = \begin{cases} \frac{\exp(g(\boldsymbol{x})_j)}{\sum_{t \in \mathcal{J}(\boldsymbol{x})} \exp(g(\boldsymbol{x})_t)} & \text{if } j \in \mathcal{J}(\boldsymbol{x}) \\ 0 & \text{if } j \notin \mathcal{J}(\boldsymbol{x}) \end{cases}, \tag{5}$$

where $g : \mathbb{R}^d \to \mathbb{R}^T$ computes the weights for each expert, and $\mathcal{J}(\boldsymbol{x})$ finds a sparse mask with at most k non-zero entries according to $\boldsymbol{x}$, i.e., $|\mathcal{J}(\boldsymbol{x})| = k, \forall \boldsymbol{x} \in \mathbb{R}^d$. For example, $\mathcal{J}(\boldsymbol{x})$ selects the indices of top-k largest logits from $g(\boldsymbol{x})$.

Below, we formally state the hypothesis space of an SMoE, which is composed of the hypotheses of both router and expert learners.

Definition 1. *Suppose all expert learners $h_1, \cdots, h_T \in \mathcal{H}$ is selected from the same hypothesis space $\mathcal{H}$, and k-sparse router function $a \in \mathcal{A}$ is chosen from the hypothesis space $\mathcal{A}$. Define the hypothesis space of the SMoE model with T experts and k-sparse router function as below:*

$$\mathcal{F}(T, k) = \left\{ f(\boldsymbol{x}) = \sum_{j=1}^{T} a(\boldsymbol{x})_j h_j(\boldsymbol{x}) : h_1, \cdots, h_T \in \mathcal{H}, a \in \mathcal{A} \right\} \tag{6}$$

2.2 Complexity Metrics and Main Proof Tools

Given a space Z and a fixed distribution D, let $S = \{z_1, \ldots, z_m\}$ be a set of examples drawn i.i.d. from D. Furthermore, let $\mathcal{F}$ be a class of functions $f : Z \to \mathbb{R}$.

Definition 2 (Rademacher complexity)**.** *The Rademacher complexity of $\mathcal{F}$ is defined as $\mathcal{R}_m(\mathcal{F}) = \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m} [\mathcal{R}_m(\mathcal{F}, \mathcal{S})]$, where $\mathcal{R}_m(\mathcal{F}, \mathcal{S}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i)\right)\right]$ is the Empirical Rademacher complexity with $\sigma_1, \ldots, \sigma_m$ to be Rademacher random variables.*

Definition 3 (Natarajan Dimension)**.** *The Natarajan Dimension of $\mathcal{H}$, denoted by $d_N(\mathcal{H})$, is the maximal size of a multiclass-shattered set $C \subset \mathcal{X}$. We say that a set $C \subset \mathcal{X}$ is shattered (Shalev-Shwartz & Ben-David, 2022) by $\mathcal{H}$ if there exist two functions $f_0, f_1 : C \to [k]$ such that For every $x \in C, f_0(x) \neq f_1(x)$. For every $B \subset C$, there exists a function $h \in \mathcal{H}$ such that $\forall x \in B, h(x) = f_0(x)$ and $\forall x \in C \backslash B, h(x) = f_1(x)$.*

Moreover, we define an Indicator function that characterizes the sparse patterns produced by the router function $a \in \mathcal{A}$.

Definition 4. *Define the following function:*

$$m(\boldsymbol{x})_j = \begin{cases} 1 & \text{if } a(\boldsymbol{x})_j \neq 0 \\ 0 & \text{if } a(\boldsymbol{x})_j = 0 \end{cases} \tag{7}$$

We use constant d_N to represent the the maximal size of a set that shatters all possible outcomes of $m(\boldsymbol{x})$.

3 Main Results

We follow the notations that we defined above and state our generalization error bound as below:

Theorem 1. *Suppose the loss function $\ell : \mathcal{Y} \times \mathbb{R} \to [0, 1]$ is C-Lipschitz, and the hypothesis space $\mathcal{F}(T, k)$ follows Definition 1, then with probability at least $1 - \delta$ over the selection of training samples, the generalization error is upper bounded by:*

$$O\left(4C\mathcal{R}_m(\mathcal{H}) + 2\sqrt{\frac{2kd_N(1 + \log(\frac{T}{k})) + d_N \log(2m) + \log(4/\delta)}{2m}}\right), \tag{8}$$

where $\mathcal{R}_m(\mathcal{H})$ is the Rademacher complexity of the expert hypothesis space $\mathcal{H}$ (cf. Definition 2), d_N is the Natarajan dimension of router function hypothesis space $\mathcal{A}$ (cf. Definition 4), m is the number of training samples, T is the total number of experts, and k is the number of selected experts.

We defer the proof to Appendix A. Below we present the applications to neural networks.

3.1 Application to Neural Networks

We highlight that our generalization bound is model-agnostic and can be applied to SMoE with different types of router or expert models, once we approximate or upper bound their corresponding complexity metrics. We now instantiate it for neural networks.

Lemma 2 (Natarajan Dimension Bound of NN with ReLU activations (Jin, 2023))**.** *Consider $\Pi_{p,S}^{ReLU}$, a neural network function class with a fixed structure S of p parameters (structure S has L layers, where the ℓ-th layer has n_ℓ nodes, $\ell \in \{1, \ldots, L\}$), then $d_N\left(\Pi_{p,S}^{ReLU}\right) \leq O(d \cdot p^2)$, where d is the number of outputs in the final layer.*

The next Rademacher Complexity Bound is useful for analyzing multilayer NNs (Bartlett et al., 2017):

Lemma 3 (Rademacher Complexity Bound for NN (Bartlett et al., 2017))**.** *Assume $\left\|\boldsymbol{x}^{(i)}\right\|_2 \leqslant c, \quad \forall i$, and let $\mathcal{H} = \left\{h_\vartheta : \|\boldsymbol{W}_i\|_{op} \leq K_i, \left\|\boldsymbol{W}_i^\top\right\|_{2,1} \leqslant b_i\right\}$, where $h_\vartheta(\boldsymbol{x}) = \boldsymbol{W}_r \phi\left(\boldsymbol{W}_{r-1} \phi\left(\boldsymbol{W}_{r-2} \ldots \phi\left(\boldsymbol{W}_1 x\right) \ldots\right)\right.$, $\|\boldsymbol{W}_i\|_{op}$ is the spectral norm of $\boldsymbol{W}_i$, $\left\|\boldsymbol{W}_i^\top\right\|_{2,1}$ is the sum of the l_2 norms of the rows of $\boldsymbol{W}_i$, and r is the number of layers. Then we have $\mathcal{R}_m(\mathcal{H}) \leq \frac{c}{\sqrt{m}} \cdot \left(\prod_{i=1}^r K_i\right) \cdot \left(\sum_{i=1}^r \frac{b_i^{2/3}}{K_i^{2/3}}\right)^{3/2}$.*

By plugging Lemma 3 and Lemma 2 in our Theorem 1, we conclude the following NN-specific statement:

Corollary 4. *Suppose the loss function $\ell : \mathcal{Y} \times \mathbb{R} \to [0, 1]$ is C-Lipschitz, and each expert is a neural network satisfying the definition in Lemma 3, and router network is a neural network satisfying the definition in Lemma 2, then with probability at least $1 - \delta$ over the selection of training samples, the generalization error is upper bounded by:*

$$O\left(4C \frac{c}{\sqrt{m}} \cdot \left(\prod_{i=1}^r K_i\right) \cdot \left(\sum_{i=1}^r \frac{b_i^{2/3}}{K_i^{2/3}}\right)^{3/2} + 2\sqrt{\frac{2kdp^2(1 + \log(\frac{T}{k})) + dp^2 \log(2m) + \log(4/\delta)}{2m}}\right) \tag{9}$$

Remark on Sparsity Awareness: We claimed that our generalization bound is *sparsity-aware*. Inspecting the generalization bound derived in Theorem 1 and Corollary 4, we remark that the term $O\left(\sqrt{k(1 + \log(\frac{T}{k}))}\right)$ is strategic and somehow explains how Sparse MoE generalizes. We notice that the term is monotonically increasing with k in the interval $k \leq T$. While the model grows larger with more available experts in parallel, it will harm the generalization by $O\left(\log(T)\right)$ factor, but we can still compensate the generalization but simply choosing less experts to counteract the problem because the error term will scale down by a factor $O(k)$.

4 Conclusion

We presented the generalization error bound for the Sparse Mixture-of-Experts model. Our bound is model-agnostic and can be applied to any SMoE structure with different router/expert base models, once we can derive their complexity metrics. More importantly, our bound explains why choosing less experts can help SMoE generalize better. Nevertheless, we acknowledge that the generalization bound can be further tightened if we introduce more complex features of SMoE such as dynamic routing. In future work, we plan to combine these features with latest deep learning theory tools to further study this intriguing SMoE generalization problem.

REFERENCES

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding mixture of experts in deep learning. *arXiv preprint arXiv:2208.02813*, 2022.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.

Ying Jin. Upper bounds on the natarajan dimensions of some function classes, 2023.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Ron Meir and Tong Zhang. *Generalization error bounds for Bayesian mixture algorithms*. Technion, Israel Institute of Technology, 2003.

Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

Huy Nguyen, Pedram Akbarian, TrungTin Nguyen, and Nhat Ho. A general theory for softmax gating multinomial logistic mixture of experts, 2023a.

Huy Nguyen, Pedram Akbarian, Fanqi Yan, and Nhat Ho. Statistical perspective of top-k sparse softmax gating mixture of experts, 2023b.

Huy Nguyen, TrungTin Nguyen, and Nhat Ho. Demystifying softmax gating function in gaussian mixture of experts, 2023c.

Huy Nguyen, TrungTin Nguyen, Khai Nguyen, and Nhat Ho. Towards convergence rates for parameter estimation in gaussian-gated mixture of experts, 2023d.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2022.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.

A APPENDIX

A.1 PROOF OF THEOREM 1

The main objective is to show the following probabilistic bound of $\sup_{f\in\mathcal{F}} |L_{\mathcal{S}}(f) - L_{\mathcal{D}}(f)|$. We first draw an i.i.d. copies of training samples: $\mathcal{S}' = \{\boldsymbol{x}'_1, \cdots, \boldsymbol{x}'_m\} \overset{i.i.d.}{\sim} \mathcal{D}^m$. Then by Lemma 7, and setting $e^{-\frac{1}{2}\epsilon^2 m} \leq 1/4$, we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |L_{\mathcal{S}}(f) - L_{\mathcal{D}}(f)| \geq \epsilon\right) \leq 2\mathbb{P}\left(\sup_{f\in\mathcal{F}} |L_{\mathcal{S}}(f) - L_{\mathcal{S}'}(f)| \geq \frac{\epsilon}{2}\right) \tag{10}$$

Our proof proceeds by reformulating the router function $a(\boldsymbol{x})$. Let us rewrite $a(\boldsymbol{x}) = \mu(\boldsymbol{x}) \odot \nu(\boldsymbol{x})$, where $\odot$ denotes the element-wise multiplication, $\mu(\boldsymbol{x}) : \mathbb{R}^d \to \{0,1\}^T$ produces a binary mask specifying the sparse expert selection ($\|\mu(\boldsymbol{x})\|_0 = k$), and $\nu(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}_+^T$ outputs the normalized weights for selected experts such that $\|a(\boldsymbol{x})\|_1 = 1$. Noting that $\nu(\boldsymbol{x})$ is dependent of the function $\mu(\boldsymbol{x})$, we define $\mathcal{V}|_\mu$ as the class of g induced by $\mathcal{A}$ and $\mu(\boldsymbol{x})$.

Now notice that $\mu(\boldsymbol{x})$ amounts to a multi-class classifier, which maps the input $\boldsymbol{x}$ to one of the sparse patterns $\mathcal{M}(2m)$. Define $\mu_1, \cdots, \mu_\Gamma$ which shatters all the possible sparse patterns produced by $2m$ data samples, then we have:

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |L_{\mathcal{S}}(f) - L_{\mathcal{S}'}(f)| \geq \frac{\epsilon}{2}\right) \tag{11}$$

$$= \mathbb{P}\left(\sup_{a\in\mathcal{A}} \sup_{\substack{h_j\in\mathcal{H},\\ \forall j\in[T]}} |L_{\mathcal{S}}(f) - L_{\mathcal{S}'}(f)| \geq \frac{\epsilon}{2}\right) \tag{12}$$

$$\leq \mathbb{P}\left(\sup_{\mu\in\{\mu_1,\cdots,\mu_\Gamma\}} \sup_{\substack{h_j\in\mathcal{H},\forall j\in[T]\\ \nu\in\mathcal{V}|_\mu}} |L_{\mathcal{S}}(f) - L_{\mathcal{S}'}(f)| \geq \frac{\epsilon}{2}\right) \tag{13}$$

$$\leq \sum_{t=1}^{\Gamma} \mathbb{P}\left(\sup_{\substack{h_j\in\mathcal{H},\forall j\in[T]\\ \nu\in\mathcal{V}|_\mu}} |L_{\mathcal{S}}(f) - L_{\mathcal{S}'}(f)| \geq \frac{\epsilon}{2} \,\middle|\, \mu = \mu_t\right) \tag{14}$$

$$\leq 2\Gamma \sup_{\mu^*\in\{\mu_1,\cdots,\mu_\Gamma\}} \mathbb{P}\left(\sup_{\substack{h_j\in\mathcal{H},\forall j\in[T]\\ v\in\mathcal{V}|_\mu}} L_{\mathcal{S}}(f) - L_{\mathcal{S}'}(f) \geq \frac{\epsilon}{2} \,\middle|\, \mu = \mu^*\right) \tag{15}$$

where Eq. 12 is obtained by Definition 1, Eq. 13 and Eq. 15 are obtained by union bound. Next, we do counting to bound Γ. Since μ is essentially a multi-class classifier with $\binom{T}{k}$ number of possible patterns, we will consider routing as a multi-class prediction problem by choosing one specific pattern among all possible patterns. By Lemma 6, we plug in the Natarajan dimension of our sparse patterns:

$$\Gamma \leq \binom{T}{k}^{2d_N} \cdot (2m)^{d_N}. \tag{16}$$

On the other hand, we bound remaining probability term by examining the expectation given a fixed masking function μ. Define function $\phi|_\mu$ conditioned on μ as:

$$\phi|_\mu(\mathcal{S}, \mathcal{S}') = \sup_{\substack{h_j\in\mathcal{H},\forall j\in[T]\\ \nu\in\mathcal{V}|_\mu}} L_{\mathcal{S}}(f) - L_{\mathcal{S}'}(f) \tag{17}$$

By McDiarmid's inequality, for any μ, we have bound:

$$\mathbb{P}\left(\phi|_\mu(\mathcal{S},\mathcal{S}') \geq \frac{\epsilon}{2}\right) \tag{18}$$

$$= \mathbb{P}\left(\phi|_\mu(\mathcal{S},\mathcal{S}') - \mathbb{E}\left[\phi|_\mu(\mathcal{S},\mathcal{S}')\right] \geq \frac{\epsilon}{2} - \mathbb{E}\left[\phi|_\mu(\mathcal{S},\mathcal{S}')\right]\right) \tag{19}$$

$$\leq \exp\left(-2m\left(\frac{\epsilon}{2} - \mathbb{E}\left[\phi|_\mu(\mathcal{S},\mathcal{S}')\right]\right)^2\right) \tag{20}$$

Combined with Eq. 10 15, 16, 20, we state, with probability at least $1-\delta$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |L_{\mathcal{S}}(f) - L_{\mathcal{D}}(f)| \geq \epsilon\right) \tag{21}$$

$$\leq 2\mathbb{E}\left[\phi|_\mu(\mathcal{S},\mathcal{S}')\right] + 2\sqrt{\frac{\log\left(\binom{T}{k}^{2d_N}(2m)^{d_N}\right) + \log(4/\delta)}{2m}} \tag{22}$$

$$\leq 2\mathbb{E}\left[\phi|_\mu(\mathcal{S},\mathcal{S}')\right] + 2\sqrt{\frac{2kd_N(1+\log(\frac{T}{k})) + d_N\log(2m) + \log(4/\delta)}{2m}} \tag{23}$$

where we used the inequality $\binom{T}{k} \leq \left(\frac{eT}{k}\right)^k$ in Eq. 23. We conclude the proof by the bounding $\mathbb{E}\left[\phi|_\mu(\mathcal{S},\mathcal{S}')\right] \leq 2C\mathcal{R}_m(\mathcal{H})$ using Lemma 5.

Lemma 5. *Consider C-Lipschitz loss function:: $\ell : \mathcal{Y}\times\mathbb{R}\to\mathbb{R}$, and follow the definition of $\phi|_\mu$ in Eq. 17, we have*

$$\mathbb{E}\left[\phi|_\mu(\mathcal{S},\mathcal{S}')\right] \leq 2C\mathcal{R}_m(\mathcal{H}) \tag{24}$$

Proof. For the sake of notation simplicity, we define a function space conditioned on a masking function μ:

$$\mathcal{F}|_\mu = \left\{ f(\boldsymbol{x}) = \sum_{j=1}^{T} \mu(\boldsymbol{x})_j \nu(\boldsymbol{x})_j h_j(\boldsymbol{x}) : h_1,\cdots,h_T \in \mathcal{H}, \nu \in \mathcal{V}|_\mu \right\} \tag{25}$$

We denote the loss function ℓ composed on $\mathcal{F}|_\mu$ as $\ell\circ\mathcal{F}|_\mu = \{\ell(f(\boldsymbol{x})) : f\in\mathcal{F}|_\mu\}$. By Lemma 8,

$$\mathbb{E}\left[\phi|_u(\mathcal{S},\mathcal{S}')\right] = \mathbb{E}\left[\sup_{\ell_f\in\ell\circ\mathcal{F}|_\mu}\left(\frac{1}{m}\sum_{i=1}^{m}\ell_f(\boldsymbol{x}_i) - \frac{1}{m}\sum_{i=1}^{m}\ell_f(\boldsymbol{x}_i')\right)\right] \tag{26}$$

$$\leq 2\mathcal{R}_m(\ell\circ\mathcal{F}|_\mu) \tag{27}$$

Since ℓ is Lipschitz function, by Lemma 9, we have

$$\mathcal{R}_m(\ell\circ\mathcal{F}|_\mu) \leq C\mathcal{R}_m(\mathcal{F}|_\mu) \tag{28}$$

Afterwards, we bound $\mathcal{R}_m(\mathcal{F}|_\mu)$ by:

$$\mathbb{E}_{\mathcal{S},\boldsymbol{\sigma}}\left[\frac{1}{m}\sup_{f\in\mathcal{F}|_\mu}\sum_{i=1}^{m}\sigma_i f(\boldsymbol{x}_i)\right] \tag{29}$$

$$= \mathbb{E}_{\mathcal{S},\boldsymbol{\sigma}}\left[\sup_{\substack{\sup_{h_j\in\mathcal{H},\forall j\in[T]}\\ v\in\mathcal{V}|_\mu}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i\sum_{j=1}^{T}\mu(\boldsymbol{x}_i)_j\nu(\boldsymbol{x}_i)_j h_j(\boldsymbol{x}_i)\right] \tag{30}$$

$$\leq \mathbb{E}_{\mathcal{S},\boldsymbol{\sigma}}\left[\sup_{\substack{h_j\in\mathcal{H},\forall j\in[T]\\ \boldsymbol{\lambda}\in\mathbb{R}_+^T,\|\boldsymbol{\lambda}\|_1=1}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i\sum_{j=1}^{T}\boldsymbol{\lambda}_j h_j(\boldsymbol{x}_i)\right] \tag{31}$$

$$= \mathcal{R}_m(\mathcal{H}), \tag{32}$$

where we notice that $\sum_{j=1}^{T}\mu(\boldsymbol{x})_j\nu(\boldsymbol{x}_j) = 1$ due to the softmax normalization over the weights of selected experts, then Eq. 31 can be relaxed by supremum over all simplex. The last equation follows from Lemma 10. Now we can conclude the proof by combining Eq. 27, 28, and 32. □

Lemma 6 (Natarajan Lemma)**.** *Given a set of finite data points $\mathcal{S}$ with $|\mathcal{S}| = m$, and a hypothesis space $\mathcal{H}$ of functions $\mathcal{S} \to [k]$ with Natarajan dimension d_N, then the growth function is bounded by:*

$$\tau_{\mathcal{H}}(m) \leq m^{d_N} \cdot k^{2d_N} \tag{33}$$

Proof. See Natarajan (1989). □

Lemma 7 (Ghost Sampling)**.** *Given $\mathcal{S}$ and $\mathcal{S}'$ with $|\mathcal{S}| = |\mathcal{S}'| = m$, we have the following inequality for any hypothesis space $\mathcal{H}$:*

$$\left(1 - 2e^{-\frac{1}{2}\epsilon^2 m}\right) \mathbb{P}\left[\sup_{h \in \mathcal{H}} |L_{\mathcal{S}}(h) - L_{\mathcal{S}'}(h)| > \epsilon\right] \tag{34}$$

$$\leq \mathbb{P}\left[\sup_{h \in \mathcal{H}} |L_{\mathcal{S}}(h) - L_{\mathcal{S}'}(h)| > \frac{\epsilon}{2}\right] \tag{35}$$

Proof. See Shalev-Shwartz & Ben-David (2022). □

Lemma 8. *Given any function class $\mathcal{F}$, for any $\mathcal{S}$ and $\mathcal{S}'$ drawn i.i.d. from $\mathcal{D}^m$ with $|\mathcal{S}| = |\mathcal{S}'| = m$, it holds that*

$$\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\sup_{f \in \mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{x}_i) - \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{x}'_i)\right)\right] \leq 2\mathcal{R}_m(\mathcal{F}) \tag{36}$$

Proof. The proof is concluded by the following derivation:

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} L_{\mathcal{S}}(f) - L_{\mathcal{S}'}(f)\right] \tag{37}$$

$$= \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\mathcal{S}'}\left[\sup_{f \in \mathcal{F}} |L_{\mathcal{S}}(f) - L_{\mathcal{S}'}(f)|\right]\right] \tag{38}$$

$$\leq \mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\mathbb{E}_{\sigma_i}\left[\sup_{f \in \mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\left(f(z_i) - \frac{1}{m}\sum_{i=1}^{m} f(z'_i)\right)\right)\right]\right] \tag{39}$$

$$\leq \mathbb{E}_{\mathcal{S},\mathcal{S}',\sigma_i}\left[\sup_{f \in \mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right) + \sup_{f \in \mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m} -\sigma_i f(z'_i)\right)\right] \tag{40}$$

$$= 2\mathcal{R}_m(\mathcal{F}), \tag{41}$$

where we introduce Rademacher random variables $\sigma_i, i = 1, \cdots, m$ in Eq. 39. □

Lemma 9. *Suppose $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$ and function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ is a C-Lipschitz function, define define $\ell \circ \mathcal{H} = \{\ell \circ h : \forall h \in \mathcal{H}\}$, then $\mathcal{R}_m(\ell \circ \mathcal{H}) \leq C\mathcal{R}_m(\mathcal{H})$.*

Proof. See Meir & Zhang (2003). □

Lemma 10. *Suppose $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$, then all functions constructed by convex combinations of $\mathcal{H}$ satisfies:*

$$\mathbb{E}_{\mathcal{S},\boldsymbol{\sigma}}\left[\sup_{\substack{h_j \in \mathcal{H}, \forall j \in [T] \\ \boldsymbol{\lambda} \in \mathbb{R}_+^T, \|\boldsymbol{\lambda}\|_1 = 1}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i \sum_{j=1}^{T}\boldsymbol{\lambda}_j h_j(\boldsymbol{x}_i)\right] = \mathcal{R}_m(\mathcal{H}). \tag{42}$$

Proof.

$$\mathbb{E}_{\mathcal{S},\boldsymbol{\sigma}}\left[\sup_{\substack{h_j\in\mathcal{H},\forall j\in[T]\\ \boldsymbol{\lambda}\in\mathbb{R}_+^T,\|\boldsymbol{\lambda}\|_1=1}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\sum_{j=1}^{T}\boldsymbol{\lambda}_j h_j(\boldsymbol{x}_i)\right] \tag{43}$$

$$=\mathbb{E}_{\mathcal{S},\boldsymbol{\sigma}}\left[\sup_{\substack{h_j\in\mathcal{H},\\ \forall j\in[T]}}\sup_{\substack{\boldsymbol{\lambda}\in\mathbb{R}_+^T,\\ \|\boldsymbol{\lambda}\|_1=1}}\frac{1}{m}\sum_{j=1}^{T}\boldsymbol{\lambda}_j\left(\sum_{i=1}^{m}\sigma_i h_j(\boldsymbol{x}_i)\right)\right] \tag{44}$$

$$=\mathbb{E}_{\mathcal{S},\boldsymbol{\sigma}}\left[\sup_{h_{j^*}\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h_{j^*}(\boldsymbol{x}_i)\right] \tag{45}$$

$$=\mathcal{R}_m(\mathcal{H}), \tag{46}$$

where Eq. 45 uses the fact that $\sum_{j=1}^{T}\boldsymbol{\lambda}_j y_j \leq \max_{j=1,\cdots,T} y_j$ for coefficients $\boldsymbol{\lambda}$. Moreover, the equality is achieved if and only if $\boldsymbol{\lambda}_j = 1$ for $j = \arg\max_{j=1,\cdots,T} y_j$ and $\boldsymbol{\lambda}_j = 0$ otherwise. □